DOCUMENT RESUME

ED 420 674 TM 028 349

AUTHOR	Reckase, Mark D.
TITLE	Analysis of Methods for Collecting Test-based Judgments.
PUB DATE	1998-04-00
NOTE	21p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Diego, CA, April 14-16, 1998).
PUB TYPE	Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	Educational Assessment; *Evaluation Methods; *Judges; Measurement Techniques; Models; Psychological Studies; *Psychometrics; Standards; *Test Use
IDENTIFIERS	National Assessment of Educational Progress; *Standard Setting

ABSTRACT

Standard setting is a fairly widespread activity in educational and psychological measurement, but there is no formal psychometric theory to guide the development of standard setting methodology. This paper presents a conceptual framework for such a psychometric theory and uses the conceptual framework to analyze a number of methods for setting standards through judges' interactions with test materials. The various standard-setting methods that have been used or are being considered for setting National Assessment of Educational Progress standards are considered. The model presented indicates that the standard is set by the agency that calls for the standard, and the task of the judges is to translate the agency's description of the standard, the task definition, to a numerical value on the reported score scale. The translation process is influenced by several features of the standard setting process, including the creation of content descriptions and the selection of the standard setting methodology. Several standard setting methods are evaluated to determine the likelihood that the judges' ratings could be used to recover the standard in a statistically unbiased way with a reasonably small standard error. Sources of variation in estimates of standards were considered, including the quality of translation of task definitions to content descriptions, the level of understanding by judges of content descriptions and item characteristics, and the amount of information acquired from the judges. Future work will emphasize formalizing concepts and developing analytic models of the standard setting process that can be used to guide data-based evaluations of the statistical quality of standards. (Contains 1 table, 1 figure, and 24 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 420 674

Analysis of Methods for Collecting Test-based Judgements

Mark D. Reckase

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☐ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
Mark Reckase
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

TM028349

This paper is prepared for the:
Annual Meeting of the National Council on Measurement in Education in San Diego, CA
April 1998

Analysis of Methods for Collecting Test-based Judgements[1]

Mark D. Reckase
Luz Bay
ACT, Inc.

Standard setting is a fairly widespread activity in the area of educational and psychological measurement. A recent issue of *Applied Measurement in Education* (Volume 11, Number 1) provides an overview of some areas where standard setting is important including licensure and certification (Plake, 1998), military training (Hanser, 1998), and the National Assessment of Educational Progress (NAEP) (Reckase, 1998). The field of educational standard setting is also well documented in such works as Jaeger's chapter in *Educational Measurement* (Jaeger, 1989), and the *Proceedings of the Joint Conference on Standard Setting for Large-scale Assessments* (NAGB & NCES, 1995). Despite the interest in the field of standard setting and the frequency of the application of standard setting methodology, there is no formal psychometric theory available to guide the development of standard setting methodology. There is a large body of very creative work concerning the development of methods and the evaluation of the outcomes of those methods [see Jaeger (1989) for a summary of some of this work], but there is no unifying theory behind those methods. This paper presents a conceptual framework for such a psychometric theory and uses the conceptual framework to analyze a number of methods for setting standards through judges interactions with test materials.

Psychometric Framework

Task definition. A standard setting study is motivated by a task definition that is provided by the agency that is responsible for setting the standards. For example, a state department of education my define the standard setting task as determining the minimum qualifications needed to be awarded a high school diploma. A professional association may define the standard setting task as determining the minimum qualifications needed to practice the profession. In this paper, the focus is on setting standards of performance on the NAEP and the policy-making agency is the National Assessment Governing Board (NAGB). The Board has defined the standard setting task by providing policy definitions for the standards to be set (NAGB, 1995).

In many respects, defining the standard setting task actually defines the standard. Indicating that the task is to determine the minimum qualifications required to receive a high school diploma indicates that a minimum level of competency is to be specified, and that it is a minimum related to the material taught in high school. The standard setting procedure

---

[1]Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, April, 1998. ©1998 by ACT, Inc. All rights reserved.

involves translating the task definition onto the numerical score scale for the test that will be used to determine whether individuals are above or below the standard. If the task definition is indicated by the symbol, *TD*, and the numerical translation of the standard onto the reported score scale for the test is indicated by γ, then the standard setting process can represented simply as

$$\gamma = f(TD). \tag{1}$$

Of course, this representation of the standard setting process vastly oversimplifies reality. There are many different types of standard setting methods and the different methods can yield notably different translations to the numerical score scale [see Shepard (1983) pp. 62-63 for a discussion of this phenomenon]. Thus, if the translation process uses method *M*, the standard setting process is more accurately represented as

$$\gamma_m = f(TD|M). \tag{2}$$

The subscript on γ acknowledges that different standard setting methods yield different standards.

Research on standard setting also suggests that the characteristics of the persons who judgementally perform the translation from task definition to numerical score scale can also affect the value of the standard that is set. For example, Busch and Jaeger (1990) indicated that public school staff translated the task definition to a numerical value that was different than that determined by college/university staff. In NAEP ALS processes that have been conducted by ACT for NAGB, results have not revealed a consistent pattern of significant differences between achievement levels cutpoints set by the three different types of judges (teacher, nonteacher educator, and general public). To the extent that differences were found, the most frequent pattern (although not the only one) was that teachers set lower cutscores than the general public judges (ACT, 1997b). Similarly, Plake, Impara & Potenza (1994) obtained mixed results. While research on the affects of the characteristics of the population of judges on the translation of the task definition to a numerical standard is ambiguous, it might be prudent to include the population of judges, *J*, as a component in the model of the standard setting process:

$$\gamma_{mj} = f(TD|M, J). \tag{3}$$

No doubt there are other features of the standard setting process that will affect the results. However, rather than complicate the representation of the function that maps the task definition to the numerical test score scale, these other features, such as time of year for the standard setting study, number of judges involved in the study, location of the study, etc. will be assumed to contribute variation to the result rather than any notable shift in the magnitude of the numerical value for the standard. The presence of this variation means that the numerical value of the standard is not really a function of the task definition, method, and population of judges—standard setting studies using the same judges, method, and task definition will likely yield different numerical values on the test score scale when other features of the study are varied. The parameter for the standard, γ_{mj}, is analogous to the true

score in classical test theory. Each replication, r, of the standard setting process provides an observed cut score on the numerical test score scale, c_{mjr}, and γ_{mj} is the expected value of the observed standard over replications:

$$c_{mjr} = f_r(TD \mid M, J) \tag{4}$$

and

$$\gamma_{mj} = E(c_{mjr}). \tag{5}$$

The variance of the distribution of estimated standards can also be computed,

$$\sigma^2_{mj} = E(c_{mjr} - \gamma_{mj})^2. \tag{6}$$

The square root of this value is a measure of the standard error of the standard. Note that although there is a standard error of the numerical standard, the reliability of the standard is not defined because the parameter for the standard (the population value) has no variation. Therefore, the classical definition of reliability as the true score variance over the observed score variance does not apply. Either there is no true score variance and the reliability is undefined, or the true score variance is zero, and the reliability is zero. In either case, the classical definition of reliability makes no sense.

Content descriptions. Up to this point, the standard setting method, *M*, has been treated as if it were a simple, single-step procedure. In actuality, a standard setting method has many different component parts and each part can be applied in a variety of different ways. For example, the task definition, *TD*, is often converted into a more specific content description, *CD*, to facilitate the translation to the test score scale. For the NAEP Achievement Levels Setting (ALS) process, these content descriptions are called Achievement Level Descriptions (ALDs). ALDs have been produced in a variety of different ways. For the 1994, 1996, and 1998 NAEP, preliminary ALDs that operationalized the NAGB policy definitions (*TD* in this context) have been developed as part of the assessment frameworks. For 1994 and 1996 NAEP ALS processes, the ALDs were modified and finalized by the judges before they were used to guide the rating process. For the 1998 NAEP ALS processes, the plan is to have a different panel finalize the ALDs (ACT, 1997a).

For NAEP, judges do not translate the task definition to the numerical scale; they translate the ALD, that is a result of a different panel's efforts, to the numerical test score scale. A different framework panel would likely produce a somewhat different ALD, resulting in a somewhat different translation to a value on the numerical score scale. Thus, the standard setting process may begin with a framework panel, f, that is asked to produce a content specific definition, $CD_f = g_f(TD)$ that is consistent with the content area that is the focus of the test. Methodology is then provided to guide judges as they translate the content specific definitions to the numerical scale, $c_{mjfr} = f_r(CD_f \mid TD, M, J)$. If the translation of the task definition to the content specific definition is replicated, as well as the full standard setting methodology, the contributions to the standard error of the cutscore can be computed for each

component. Brennan (1995) has suggested this type of partitioning of the error variance in a standard setting study through the use of generalizability theory.

The judges' tasks. A judge in a standard setting study has two basic tasks. First, the judge must comprehend the content description for the standard and create for themselves an internal representation of the skills and knowledge that a person that matches the content description would have. The second task is to interact with the test materials or examinee population to generate information that can be used to compute an estimate of their translation of the content description to the numerical test score scale. If a judge performs these two tasks perfectly, that is if they completely understand the content distribution and if they can apply that understanding to the standard setting task in an accurate and consistent way, then the result is an error free translation of the content description to the numerical test score scale.

Using the conceptual model that has been developed so far, the error free standard from a judge is given by the function

$$\gamma_{jmf} = f(j|CD_f, TD, M). \tag{7}$$

This is not the only form that can be used to represent the relationship between the standard and the standard setting process. If it is believed that thorough training and discussion of the content descriptions, *CD*, will result in a common internal representation of a person who meets the standard for all judges, and if it is believed that with sufficient training all judges can perform the standard setting process accurately, then there is no need to have an index for judge on the standard. The error free standard should be the same for every judge that has been properly informed and trained,

$$\gamma_{mf} = f(CD_f, TD, M). \tag{8}$$

Alternatively, if it is believed that the makeup of the group of judges can influence the translation of the standards to the numerical scale, and if the characteristics of the judge will influence the internal representation of the content description, then both the group of judges, *J*, and the interpretation of the content description by the *j*th judge, $I_j(CD_f)$, need to be included in the model,

$$\gamma_{jmf} = f(j|I_j(CD_f), CD_f, TD, M, J). \tag{9}$$

Depending on the characteristics of the reported score for the test, the γ-metric may be the true score metric, or an IRT based θ-metric, or some other function of examinees performance on the test. For the application considered here, standard setting on NAEP, the θ-metric is appropriate since IRT methodology is used to define the reporting score scale for NAEP. If the judges fully comprehend the content descriptions, and if they apply the standard setting methodology without error, each judge will have an error free cut score, $\theta_{jmf} = \gamma_{jmf}$. The methodology of standard setting has the purpose of collecting information from the judges that can be used to estimate the value of θ_{jmf}.

If training and discussion can bring all of the judges to a common understanding of the content description and a high level of competence with the standard setting methodology, then the value of θ_{jmf} will be equal for all judges and each judge can be considered as a replication. If training and discussion are considered to be insufficient to reach common understanding and a high level of competence, than each judge will be expected to have his or her own value of θ_{jmf} depending on his or her interpretation of the content description and approach to the standard setting methodology. The results under the latter conditions are many individual standards. Because a single standard is usually desired, a method must be devised to determine the actual standard to be used. Each judge's standard is equally good on statistical grounds; other factors, such as the quality of a judge's participation in the process, the judge's knowledge of the examinee population, or the desire to set a low or high standard, need to be brought to bear to determine the final single value to be used.

Figure 1 provides an example of the distribution of standards implied by judges' ratings on the NAEP Science Test. Each letter in the distribution shows the location of the estimated standard for one judge. If the judges are considered as replicates, the mean of the distribution is likely to be a good estimator of θ_{mf}. If each judge provides a unique interpretation of the *TD* and *CD*, than more judgement by the policy-making agency is needed to select a value for θ. For example, a liberal standard could set by selecting the lowest θ-estimate from all of the judges. Selection of the lowest value implies that the judges' estimates are not related to the single standard implied by *TD*, but that the opinions of the judges drive the standards, and any one opinion is as good as another one. The difference in philosophical approach to standard setting can result in quite different standards, as the difference between the mean of the distribution (171) in Figure 1 and the lowest score (164) shows.

Insert Figure 1 about here

Sources of Variation in Standard Setting Studies

This conceptual framework provides a means for considering factors that will likely result in differences in standards resulting from standard setting studies. These will be summarized here for further consideration.

(1) *Translating task definitions to content descriptions.* If two equivalent groups of individuals were assigned to independently translate the task definition to a content description, the results will not likely be the same. Further, the magnitude of the differences will likely depend on how seriously the individuals approach the task and how much effort is dedicated to it. A haphazard approach to the task will likely yield highly variable results. In

most cases, this translation is done only once so it is not possible to determine the effects on the outcome of the standard setting study.

(2) *Judges' interpretations of content descriptions, or task definitions.* Reading content descriptions or task definitions may conjure quite different internal representations of skills and knowledge among judges. These representations can be made more similar through discussion and elaboration of the content descriptions. To the extent that developing a common frame of reference is given priority in a standard setting study, variation due to differences in interpretation can be minimized. If this source of variation is ignored, either by not producing a content description, or by using an ambiguous one, or through inadequate time and effort to reach common understanding, the result could be large variation in the internal standards developed by the judges.

(3) *Understanding of the standard setting process.* As part of the standard setting process, judges are asked to interact with test materials, examinees responses, or the examinees themselves to provide information that can be used to determine each judge's translation of the content description of task definition to the numerical score scale for a test. Lack of a clear understanding of the translation task will lead to variation in the application of the standard setting process. This will also lead to variation in the judges' standards.

An Analysis of Standard Setting Methods

As indicated by Jaeger (1989), Plake (1998), Hanser (1998) and others, there are a wide variety of methods used for the translation of a task definition to a numerical value on a test's score scale. These methods can be analyzed using the conceptual framework provided above to infer the cognitive processes that are required to apply the methodology. The statistical underpinnings of the method can also be listed. For purposes of demonstration, an analysis of the modified Angoff (1971) method will be presented first, then other methods will be considered.

The modified-Angoff method. The modified-Angoff method (Angoff, 1971; p. 515) requires that judges first gain an understanding of the content description or task definition that guides the standard setting process. From that understanding, the judges develop an internal representation of the least able examinee that exceeds the standard defined by the content description. In the NAEP context, the least qualified examinee for an Achievement Level implies a point on the θ-scale that is a judge's cut score. The goal of the standard setting process is to estimate each judge's cut score on the θ-scale.

If the judge's θ-value were known, and if the item characteristic curve (ICC) for a test item were known as well, then the probability of correct response on the test item that corresponds to the judge's θ-value can be computed directly from the IRT function. For

example, if the IRT function is given by

$$P(\theta_j) = P(u_{ij}=1 \mid \theta_j, a_i, b_i, c_i) = c_i + (1-c_i)\frac{e^{a_i(\theta_j - b_i)}}{1+e^{a_i(\theta_j - b_i)}}, \quad (10)$$

where θ_j is the value on the θ-scale,
a_i is the discrimination parameter for the item,
b_i is the difficulty parameter for the item,
and c_i is the lower asymptote (guessing) parameter for the item,

the probability of correct response to the item for the minimally competent examinee from judge *j*'s perspective is computed by inserting the judge's θ_{jmf} for θ_j and evaluating Equation 10. If there are 100 examinees exactly at θ_j, the number that would be expected to get the item correct would be $P(\theta_j)$ x 100.

Unfortunately, θ_j is the parameter that is to be estimated—it is unknown—and the parameters of the ICC are typically estimated from examinee responses to the test items, not from judges' ratings. The information that is obtained from the judges when the Angoff procedure is used is each judge's estimate of $P(\theta_{mjf})$ for each item. This estimate is obtained by either asking for the probability directly, or by asking the judges to provide the number of examinees out of 100 that are exactly at the cut score that will get the test item correct. Of course, the estimates contain error that is related to the preciseness of the judge's internal representation of the content description and his or her understanding of the functioning of the test item.

Once the estimates of $P(\theta_{jmf})$ have been obtained, there are a number of approaches to using the information to estimate θ_{jmf}. The most straight forward approach is to assume that the judge is using the same ICC that was estimated from the examinees' responses and treat $P(\theta_j)$ and the item parameters as known and solve for θ_j in Equation 10. This results in a distinct θ-estimate for the judgement of each item. The full collection of these estimates for θ_{jmf} provide an estimate of the error distribution for the judge's estimates. The mean of the distribution can be used as an estimate of θ_{jmf} and the standard deviation of the distribution is an estimate of the standard error of the judge's estimate. This standard error only considers the variance associated with interpretations of the content descriptions and misjudgments of the characteristics of the ICCs for the items.

Other methods have been used to estimate θ_{jmf} from judges' ratings generated from the Angoff process. The probability estimates from all of the items can be summed to provide an estimate of the true score on the set of items. The θ-value that corresponds to the true score estimate can be obtained from the test characteristic curve for the set of items. In this case, only a single estimate is produced for the θ_{jmf} for a judge so no estimate of the standard error of the translation to the θ-scale is available. Bayesian and maximum likelihood procedures have also been developed (Davey, Fan, & Reckase, 1996). In general, Davey et al. (1996) found that maximum likelihood and Bayesian procedures for estimating θ_{jmf} resulted in smaller standard errors than the simpler mapping procedures because they take into account

the varying characteristics of the test items. Both of these methods have been used by ACT to estimate Achievement Levels on NAEP.

In order to accurately estimate a judge's θ_{jmf} value, the judge must provide estimates of $P(\theta_j)$ that are statistically unbiased. That is, the estimates should be no more likely to be above the value specified by the IRT model than below that value. The mean of the sampling distribution of the estimates provided by the judge should approach $P(\theta_j)$ as the number of judgements increases. For this to occur in practice, judges need to have a clear understanding of the connection between the content descriptions and the item characteristics, and they need to know the form of the ICC for the item.

A well trained judge will likely understand the connection of the content description to the test item, but it is unlikely that they will have a good sense of the form of the ICC during the initial estimation of $P(\theta_j)$. The modified Angoff procedure typically provides information about item difficulty as feedback after a first round of ratings. This would help to give an ordering by *b*-parameter for the items, but not other characteristics. An understandable representation of the ICCs should be provided.

The modified-Angoff procedure also provides feedback about the relative position of the standards set by the judges. This feedback serves to let judges know if they are extreme in their estimates of the probability of correct response. The fact that this information is provided suggests that the model given in Equation 8 is the basis for the procedure. If the procedure were functioning without error, all judges would be expected to arrive at the same standard. But, because of lack of knowledge of the ICCs for the items, and differences in interpretation of the content description, there is a distribution of standards from the judges rather than a single value. But, because Equation 8 is the model for the procedure, it makes sense to use the mean of the distribution of standards as an estimate of the target standard for the group.

If the modified-Angoff procedure is working well, it is clear that the estimates of $P(\theta_j)$ can be used to recover θ_{jmf}. The value of the standard is derived from the ICC and the value of the probability. This is an important property of the Angoff procedure. Other procedures may not provide a means for recovering the standard that underlies a judges ratings.

Other Standard Setting Procedures

In the course of the work that ACT has done for NAGB, a wide variety of standard setting procedures have been implemented and evaluated. Table 1 provides a list of the procedures and indicates the circumstances under which they were applied. The time limitations of an NCME presentation prevent analysis of all of these procedures. But a few of them will be subjected to the same type of analysis applied to the modified Angoff procedure to demonstrate a basic framework for analyzing other standard setting procedures.

Insert Table 1 about here

Item Score String Estimation (ISSE). The ISSE method was proposed for the 1998 NAEP ALS processes for civics and writing. Two field trials have been implemented (one for each subject) where judges used the ISSE in rating a mix of dichotomously and polytomously scored items in one field trial (Bay, 1998a), and rating only polytomous items in another field trial (Bay, 1998b). The purpose of the field trials was to compare [illegible] to the "mean estimation" method (see Table 1). For the ISSE method, judges are asked to determine the most likely score for an item for an examinee exactly at the cut score, rather than the probability of correct response as required by the modified Angoff procedure. The most likely score is estimated for each item on a test, resulting in a item score string that is similar to the string of item scores produced by an examinee. From that string, a θ-value can be estimated from judges' ratings using the same procedures that are used to estimate examinees' θ-values. From a strictly statistical perspective, if a dichotomously scored test item has a probability of correct response greater than .5, than a correct response is more likely than an incorrect response and the judge should indicate that a 1 is the more likely score. If the probability of correct response is less than .5, than the more likely score should be 0.

For a judge to accurately provide information for this method, he or she has to have a clear understanding of the connection between the content description and the item and at least be able to tell whether the probability of correct response to the item given their standard is greater or less than .5. This procedure was developed as an easier alternative to the modified-Angoff procedure. It is interesting to note that Angoff (1971, p. 514) suggested basically the same procedure. The modified-Angoff procedure was a variation mentioned in a footnote to providing estimated item scores for each item. The Angoff item score procedure does not require accurate estimation of probabilities.

A critical question about the ISSE procedure is whether judges can provide a score string that will recover the value of θ that is their standard. In most cases, the answer is no. A simple example can be used to show the problem with the procedure. Suppose a dichotomously scored test is composed of 50 questions that all have exactly the same ICCs. Also suppose that at the judge's value of θ_{jmf} the probability of correct response is .8 for each item. The most likely response for each item is 1 and the response string for the judge would be all 1s. If a maximum likelihood estimation procedure is used to estimate θ, the resulting estimate is positive infinity rather than the finite value of the judge's cutscore. The reason that θ_{jmf} is not well estimated is that while each item has a most likely score of 1, the most likely number of 1s for the 50 item test is 40. Thus, the most likely response string would have 10 0-values in it. In general, if the most of the items have probabilities greater than .5 at the judge's standard, the ISSE will result in a estimate of the standard with a positive bias. If most items have a probability of less than .5, the result is an estimate of the cut score with a negative bias.

To recover the judge's cut score, the ISSE procedure would have to require the judge to first give an estimate of the total number of 1s in the response string, and then indicate which items have scores of 1 or 0. This would seem to be extremely difficult for a judge to do. While the ISSE procedure appears easier to apply on the surface, in reality, the reduced accuracy of rating that is required results in statistically biased estimates of the standards.

Item mapping. As part of a series of studies to support the 1998 NAEP ALS processes, field trials using item mapping procedures will be implemented (ACT, 1997a). The item mapping procedure that will be used is a variation of the "Bookmark" procedure by Lewis, Mitzel, and Green (1996). The analysis presented here will be of the original procedure described by Lewis et al. (1996). To simplify this analysis, the bookmark procedure for dichotomous items will be considered. Lewis et al (1996) describe how the bookmark procedure can be applied to polytomously scored items.

As with the other procedures, the bookmark approach starts with a task definition and content description, and the judges are assumed to have an ideal, error-free cut score that is consistent with their interpretation of the content description. The information provided by a judge that is used to estimate the location of θ_{jmf} is his or her reaction to an ordered set of test items. Test items are ordered on a θ-scale using the θ-value for the item that yields a probability of correct response of 2/3. The judge reviews the ordered list of test items and determines which two items are closest to yielding a probability of correct response of 2/3 at his or her value of θ_{jmf}, one slightly above the cutscore, and the other slightly below the cutscore. The estimate of θ_{jmf} is the average θ-value for the two items based on the probability of 2/3.

As with the other procedures, the bookmark procedure requires that the judges have a good understanding of the content descriptions and the relationship to the test items. They also need to have some sense of the ICCs for the items because they need to be able to determine whether an item has close to a 2/3 probability of a correct response near the θ_{jmf} value. If they can do the task, the method should recover the judge's cut score. Thus, the bookmark and Angoff procedure should give similar results. The differences in the two procedures is that the bookmark procedure uses much less information. Only θ-estimates for two items are used to estimate the standard. This may result in a procedure that is easy for judges to apply, but it may also yield estimates of standards with large standard errors. The modified-Angoff procedure provides an estimate of a judge's cut score for every item in a test, resulting in an estimate of the cut score that is the average of many data points. This should result in a procedure with a smaller standard error of estimate than the bookmark procedure. To gain more data points, the bookmark procedure could be performed on multiple subsets of the full test item pool and the results from each subset averaged. This variation in the procedures would allow an estimate of the standard error of the bookmark estimates to be obtained.

Paper selection. To demonstrate the analysis of a standard setting procedure for polytomously scored test items, the paper selection method will be considered. The paper

selection procedure was used in the 1992 ALS processes for reading, writing, and mathematics (ACT, 1993). As for the other procedures, the paper selection method requires that the judges have a thorough understanding of the content descriptions. The judges then conceptualize the least able person that meets the standard. This person has an IRT scale value of θ_{jmf} as was the case for the other procedures. The judges are next presented with a set of responses to a polytomous item that span the range of possible responses. The judges' task is to select a number of the papers that are as similar as possible to the paper that would be produced by a person at θ_{jmf}. The judges are presented the papers without scores, but they are well informed about the scoring rubric that is used to evaluate the papers. The data that is produced from this process is one or more papers for each item that is presented to a judge. These papers have been scored prior to the selection process, so the procedure also produces a score string for a person at the cut score.

The score string produced by a judge and the item parameters for the item can be used with the same estimation procedure that is used to score an examinees responses. For NAEP, these procedures are based on the generalized partial credit IRT model (Muraki, 1992). This model is given by

$$P_{jk}(\theta) = \frac{e^{\sum_{v=1}^{k} a_j(\theta - b_{jv})}}{\sum_{c=1}^{m_j} e^{\sum_{v=1}^{c} a_j(\theta - b_{jv})}}, \qquad (11)$$

where $P_{jk}(\theta)$ is the probability of response k to item j,
a_j is an item discrimination parameter,
and b_{jv} is an item step parameter.

The expected score on an item for a particular θ-value is given by

$$S_j(\theta) = \sum_{k=1}^{m_j} kP_{jk}(\theta). \qquad (12)$$

Note that the expected score is a continuous variable even though the scoring of the item is discrete. If the score from a single paper is used to estimate θ_{jmf} for a judge, only a limited number of θ-values are possible because of the discrete nature of item scores. However, if the judge selects multiple papers as representing the performance of an examinee at the cut score, then the scores for those papers can be averaged, resulting in more options for θ-values. The "mean estimation method" (see Table 1) asks judges to directly estimate the mean score for an item, allowing all possible values of θ.

For a good estimate of the cut score to be recovered using the paper selection method, the judges need to have a good understanding of the content descriptions and how they relate to examinees papers. Further, the judges need to be well versed in the use of the scoring rubrics. It would be optimal if the judges were formally trained to score the papers to the level of performance of the actual scorers. A lower level of scoring accuracy will likely lead

to inaccurate selection of papers. The selections may be statistically biased if misunderstandings about the scoring process result in either over estimation or underestimation of scores.

The score string for a judge's selection of papers is produced by other scorers. The scores on the papers are not perfectly accurate as well. That means that even if a judge picks appropriate papers, the score that is assigned might not be consistent with the criteria used for selection. These features of the paper-selection method suggest that, while it may provide unbiased recovery of the cut score, the standard error of estimate for the cut score is likely to be fairly large.

Summary and Discussion

A general psychometric model for the standard setting process has been described in this paper and various standard setting methods have been used or are being considered for a NAEP ALS process have been logically analyzed using the model as a framework. The model indicates that the standard is set by the agency that calls for the standard and the task of the judges is to translate the agencies description of the standard, the task definition, to a numerical value on the reported score scale.

The translation process is influenced by several features of the standard setting process. One is the creation of content descriptions to operationalize the task description for a particular content area. Another is the selection of the standard setting methodology. Given a task definition, content description, and method, a true standard, analogous to a true score for an examinee, is considered to exist, at least conceptually. Standard setting methods differ in their ability to recover the true standard. Several standard setting methods were evaluated to determine the likelihood that the judges' ratings could be used to recover the standard in a statistically unbiased way with a reasonably small standard error.

Sources of variation in estimates of standards were considered including the quality of translation of task definitions to content descriptions, the level of understanding for judges of the content descriptions and the characteristics of the items, and the amount of information that is acquired from the judges.

This paper is a first attempt at developing a psychometric theory of standard setting. Future work will emphasize formalizing concepts and developing analytic models of the standard setting process that can be used to guide data-based evaluations of the statistical quality of standards.

References

ACT, Inc. (1993). *Setting achievement levels on the 1992 National Assessment of Educational Progress in Mathematics, Reading, and Writing: A technical report on reliability and validity.* Iowa City, IA: Author.

ACT, Inc. (1994). *Design document: setting achievement levels on the 1994 National Assessment of Educational Progress in Geography and U.S. History and the 1996 National Assessment of Educational Progress in Science.* Iowa City, IA: Author

ACT, Inc. (1997a). *Developing achievement levels on the 1998 NAEP in Civics and Writing: Design document.* Iowa City, IA: Author.

ACT, Inc. (1997b). *Setting achievement levels on the 1996 National Assessment of Educational Progress in Science: Final report.* Iowa City, IA: Author.

Anderson, S. B. & Helmick, J. S. (Eds.). (1983). *On educational testing.* San Francisco: Jossey-Bass.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational Measurement (2nd Edition).* Washington DC: American Council on Education.

Bay, L. (1998a, March). *1998 NAEP achievement levels-setting process field trial 1 for civics.* A report prepared for the meeting of the 1998 NAEP Achievement Levels-setting Project Technical Advisory Committee for Standard Setting (TACSS), Chicago.

Bay, L. (1998b, March). *1998 NAEP Writing achievement levels-setting process field trial 1: ISSE vs. ME.* A report prepared for the meeting of the 1998 NAEP Achievement Levels-setting Project Technical Advisory Committee for Standard Setting (TACSS), Chicago.

Bay, L., Chen, L. & Reckase, M. (1997, October). *The grid: a possible method for the 1998 NAEP Writing Achievement Levels-setting Process.* A report prepared for the meeting of the 1998 Achievement Levels-setting Project Technical Advisory Committee for Standard Setting (TACSS), St. Louis.

Brennan, R. L. (1995). Standard setting from the perspective of generalizability theory. *Proceedings of the Joint Conference on Standard Setting for Large-scale Assessments of the National Assessment Governing Board (NAGB) and the National Center for Educational Statistics (NCES).* Washington, DC: U.S. Government Printing Office.

Busch, J. C. & Jaeger, R. M. (1990). Influence of type of judge, normative information, and discussion on standards recommended for the National Teacher Examination. *Journal of Educational Measurement, 27,* 145-163.

Davey, T., Fan, M., & Reckase, M. D. (1996, April). *Some new methods for mapping ratings to the NAEP theta scale to support estimation of NAEP achievement level boundaries.* Paper presented at the annual meeting of the National Council on Measurement in Education, New York.

Hanser, L. M. (1998). Lessons for the National Assessment of Educational Progress from military standard setting. *Applied Measurement in Education, 11*(1), 81-95.

Hofstee, W. K. B. (1983). The case for compromise in educational selection and grading. In S. B. Anderson & J. S. Helmick (Eds.), *On educational testing.* San Francisco: Jossey-Bass, 109-127.

Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.). New York: American Council on Education & Macmillan, 485-514.

Lewis, D. M., Mitzel, H. C., & Green, D. R. (1996, June). *Standard setting: A bookmark approach.* Paper presented at the CCSSO National Conference on Large Scale Assessment, Phoenix.

Linn, R. L. (Ed.) (1989). *Educational measurement* (3rd ed.). New York: American Council on Education & Macmillan.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*(2), 159-176.

NAGB (1995). *Developing student performance levels for the National Assessment of Educational Progress: policy statement.* Washington, DC: Author.

NAGB & NCES (1995). *Proceedings: Joint conference on standard setting for large-scale assessments* (Vol. 2). Washington, DC: U.S. Government Printing Office.

Plake, B. S. (1998). Setting performance standards for professional licensure and certification. *Applied Measurement in Education, 11*(1), 65-80.

Plake, B. S., Impara, J. C. & Potenza, M. T. (1994). Content specificity of expert judgments in a standard-setting study. *Journal of Educational Measurement, 31*(4), 339-347.

Reckase, M. D. (1998). Converting boundaries between National Assessment Governing Board performance categories to points on the National Assessment of Educational Progress score scale: The 1996 Science NAEP process. *Applied Measurement in Education, 11*(1), 9-21.

Shepard, L. A. (1983). Standards for placement and certification. In S. B. Anderson & J. S. Helmick (Eds.), *On educational testing*. San Francisco: Jossey-Bass, 61-90.

Table 1

Checklist of Rating Methods Used in NAEP Achievement Levels Setting Processes

Rating Method	92 Reading		92 Writing ALS	92 Math ALS	94 Geography		94 U.S. History		96 Science				98 Civics		98 Writing		
	PS	ALS			PS	ALS	PS	ALS	PS1	PS2	ALS	Rec	FT1	FT2	FT1	FT2	FT3
Modified Angoff[a] (MA) (ACT, 1994)	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓						
Paper selection[b] (PS) (ACT, 1994)	✓	✓	✓	✓													
Mean estimation[b] (ME) (ACT, 1994)					✓	✓		✓	✓	✓	✓				✓		
Estimated score point percentage[b] (EP) (ACT, 1994)					✓		✓										
Modified percentage estimate[b] (MP) (ACT, 1994)							✓										
Hybrid method[b] (H) (ACT, 1994)					✓		✓										
Item score string estimation[c] (ISSE) (ACT, 1997a)													✓		✓		
The grid[d] (G) (Bay, Chen, Reckase, 1997)																?	
Item mapping[c] (IM) (ACT, 1997a)												✓		✓			✓

Notes

ALS: Achievement Levels Setting Study -- when achievement levels recommended to NAGB are set.

PS: Pilot Study -- "dress" rehearsal for the ALS.

PS*n*: Pilot study number *n*.

FT*n*: Field Trial number *n* -- when new rating methods are tried for the first time for the 1998 NAEP ALS process.

Rec: The recommended achievement levels for grade 8 science NAEP were results of the reconvening of the grade 8 panel.

---

[a]Used for dichotomous items only.
[b]Used for polytomous items only.
[c]Used for both dichotomous and polytomous items.
[d]Judges rate two items (i.e., writing prompts) at a time.

Figure 1

Science Achievement Levels Setting: Grade 12 Round 1
Rater Location Feedback: Proficient Level (No. of Raters = 32)